DETAILED ACTION

Claim Status

1. Claims 1, 5-7, 9-12, and 31 are pending.

Examiner's Amendment

2. An examiner's amendment to the record appears below. Should the changes and/or additions be unacceptable to applicant, an amendment may be filed as provided by 37 CFR 1.312. To ensure consideration of such an amendment, it MUST be submitted no later than the payment of the issue fee.

3. Authorization for an examiner's amendment was given in a telephone interview with Benjamin J. Hauptman (reg. 29, 310) on October 24, 2008.

4. **In the claims:**

Claims 1, 5-7, 9-12, and 31 have been amended. Please replace all prior claims with the claims below.

Claim 1 :

A computer implemented method of clustering documents, each clustering document having one or plural document segments in an input document, said method comprising steps:

(a) obtaining a co-occurrence matrix for the input document by using a computer, the co-occurrence matrix is a matrix reflecting occurrence frequencies of terms and co-occurrence frequencies of term pairs, and obtaining an input document frequency matrix for a set of input

documents based on occurrence frequencies of terms or term pairs appearing in the set of input documents wherein said step (a) further includes:

generating an input document segment vector for each input document segment of said input document segments based on occurrence frequencies of terms appearing in said each input document segment;

obtaining the co-occurrence matrix for the input document from input document segment vectors; and

obtaining the input document frequency matrix from the co-occurrence matrix for each document;

(b) selecting a seed document from a set of remaining documents that are not included in any cluster existing, and constructing a current cluster of an initial state based on the seed document, wherein said selecting and said constructing comprises:

constructing a remaining document common co-occurrence matrix for the set of the remaining documents based on a product of corresponding components of co-occurrence matrices of all documents in the set of remaining documents; and

obtaining a document commonality of each remaining document to the set of the remaining documents based on a product sum between every component of the co-occurrence matrix of each remaining document and the corresponding component of the remaining document common co-occurrence matrix;

extracting a document having highest document commonality to the set of the remaining documents; and

constructing initial cluster by including the seed document and neighbor documents similar to the seed document;

(c) making documents, which have document commonality to a current cluster higher than a threshold, belong temporarily to the current cluster;

wherein said making comprising:

constructing a current cluster common co-occurrence matrix for the current cluster and a current cluster document frequency matrix of the current cluster based on occurrence frequencies of terms or term pairs appearing in the documents of the current cluster;

obtaining a distinctiveness value of each term and each term pair for the current cluster by comparing the input document frequency matrix with the current cluster document frequency matrix;

obtaining weights of each term and each term pair from the distinctiveness values;

obtaining a document commonality to the current cluster for each document in a input document set based on a product sum between every component of the co-occurrence matrix of the input document and the corresponding component of the current cluster common co-occurrence matrix while applying the weights to said components; and

making the documents having document commonality to the current cluster higher than the threshold belong temporarily to the current cluster;

(d) repeating step (c) until number of documents temporarily belonging to the current cluster does not increase;

(e) repeating steps (b) through (d) until a given convergence condition is satisfied; and

(f) deciding, on a basis of the document commonality of each document to each cluster, a cluster to which each document belongs and outputting said cluster.

Claim 2 (canceled)

Claim 5 :

The clustering method according to claim 1, wherein the convergence condition in said step (e) is satisfied when

(i) the number of documents whose document commonalities to any current clusters are less than a threshold becomes 0, or

(ii) the number is less than a threshold and does not increase.

Claim 6 :

The clustering method according to claim 1, wherein said step (f) further includes: checking existence of a redundant cluster, and removing, when the redundant cluster exists, the redundant cluster and again deciding the cluster to which each document belongs.

Claim 7 :

A computer implemented method of clustering documents each having one or plural document segments in an input document, said method comprising steps:

(a) using a computer to obtain a co-occurrence matrix for the input document, obtaining a co-occurrence matrix S^r for a input document D_r based on occurrence frequencies of terms or term pairs appearing in the set of input documents;

wherein in step (a), each mn component S^r_{mn} of the co-occurrence matrix S^r of the document D_r is determined in accordance with:

$$S^r_{mn} = \sum_{y=1}^{Y_r} d_{rym} d_{ryn}$$

where:

m and n denote m^{th} and n^{th} terms, respectively, among M terms appearing in the set of input documents,

D_r is r^{th} document in a document set D consisting of R documents;

Y_r is number of document segments in the document D_r, wherein said d_{rym} and d_{ryn} denote existence or absence of the m^{th} and n^{th} terms, respectively, in y^{th} document segment of the document D_r, and

S^r_{mm} represents number of document segments in which the m^{th} term occurs and S^r_{mn} represents co-occurrence counts of document segments in which the m^{th} and n^{th} terms co-occur;

(b) selecting a seed document from a set of remaining documents that are not included in any cluster existing, and constructing a current cluster of an initial state based on the seed document, wherein said selecting and said constructing comprise:

constructing a remaining document common co-occurrence matrix T^A for a set of the remaining documents based on co-occurrence matrices of all documents in the set of remaining documents;

obtaining a document commonality of each remaining document to the set of the remaining documents based on the co-occurrence matrix S^r of each remaining document and the remaining document common co-occurrence matrix T^A;

extracting the document having a highest document commonality to the set of the remaining documents; and

constructing a initial cluster by including the seed document and neighbor documents similar to the seed document;

(c) making documents having document commonality higher than a threshold belong temporarily to the current cluster;

(d) repeating step (c) until a number of documents temporarily belonging to the current cluster does not increase;

(e) repeating steps (b) through (d) until a given convergence condition is satisfied; and

(f) deciding, on basis of the document commonality of each document to each cluster, a cluster to which each document belongs and outputting said cluster.

Claim 9 :

The method according to claim 7, wherein in step (b), the remaining document common co-occurrence matrix T^A is determined on the basis of a matrix T;

wherein the matrix T has an mn component determined by

$T_{mn} = \prod_{r=1}^{R} S^r_{mn}$ and

$S^r_{mn} > 0$

the matrix T^A has an mn component determined by

$T^A_{mn} = T_{mn}$ when $U_{mn} > A$,

$T^A_{mn} = 0$ otherwise,

where

U_{mn} represents an mn component of a document frequency matrix of the set of remaining documents wherein U_{mm} denotes the number of remaining documents in which the m^{th} term occurs and U_{mn} denotes the number of remaining documents in which the m^{th} and n^{th} terms co-occur; and

A denotes a predetermined threshold.

Claim 10:

The method according to claim 9, further comprising:

determining a modified common co-occurrence matrix Q^A on the basis of T^A ; and in step (b), obtaining the document commonality of each remaining document to the set of the remaining documents based on the co-occurrence matrix Sr of each remaining document and the modified common co-occurrence matrix Q^A;

the matrix Q^A having an mn component determined by

$Q^A_{mn} = \log T^A_{mn}$ when $T^A_{mn} > 1$,

$Q^A_{mn} = 0$ otherwise.

Claim 11 :

The method according to claim 10, wherein in step (b),

the document commonality of each remaining document P having a co-occurrence matrix S^P with respect to the set of remaining documents is given by

$$com_q(D', P; Q^A) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{M} Q_{mn}^{A} S_{mn}^{P}}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M} (Q_{mn}^{A})^2}\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M} (S_{mn}^{P})^2}}.$$

Claim 12 :

The method according to claim 10, wherein in step (b), the document commonality of each remaining document P having a co-occurrence matrix S^P with respect to the set of remaining documents is given by

$$com_q(D', P; Q^A) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{M} T_{mn}^{A} S_{mn}^{P}}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M} (T_{mn}^{A})^2}\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M} (S_{mn}^{P})^2}}.$$

Claims 23-24 (canceled)

Claims 27-28 (canceled)

Claims 29-30 (canceled)

Claim 31 :

The method according to claim 1, wherein the remaining document common co-occurrence matrix or the current cluster common co-occurrence matrix reflects co-occurrence frequencies at which pairs of different terms co-occur in each document of the remaining documents or the current cluster.

Allowable Subject Matter

5. Claims 1, 5-7, 9-12, and 31 are allowed.

6.. The following is a statement of reasons for the indication of allowable subject matter.

The claims are directed to clustering documents each having one or plural document segments in an input document set. The present invention allows for increased accuracy for clustering documents. In doing so the system detects a seed document, calculates document set commonalities, and detects terms and term pairs not distinctive to a particular cluster. The seed document is utilized as a basis for starting a cluster. Clusters are based on term similarity. The invention expands the seed document, hence the clusters are formed. In doing so, the invention determines terms and term pairs not distinct to a particular cluster. The non-common terms are determined, excluded from current cluster, and thus deciding, on basis of the document commonality of each document to each cluster, a cluster to which each document belongs.

With respect to the independent claim 1, the prior art of record, single or in combination, does not teach or fairly suggest the step of: "(a) obtaining a co-occurrence matrix for the input

document by using a computer, the co-occurrence matrix is a matrix reflecting occurrence frequencies of terms and co-occurrence frequencies of term pairs, and obtaining an input document frequency matrix for a set of input documents based on occurrence frequencies of terms or term pairs appearing in the set of input documents wherein said step (a) further includes: generating an input document segment vector for each input document segment of said input document segments based on occurrence frequencies of terms appearing in said each input document segment; obtaining the co-occurrence matrix for the input document from input document segment vectors; and obtaining the input document frequency matrix from the co-occurrence matrix for each document; (b) selecting a seed document from a set of remaining documents that are not included in any cluster existing, and constructing a current cluster of an initial state based on the seed document, wherein said selecting and said constructing comprises: constructing a remaining document common co-occurrence matrix for the set of the remaining documents based on a product of corresponding components of co-occurrence matrices of all documents in the set of remaining documents; and obtaining a document commonality of each remaining document to the set of the remaining documents based on a product sum between every component of the co-occurrence matrix of each remaining document and the corresponding component of the remaining document common co-occurrence matrix; extracting a document having highest document commonality to the set of the remaining documents; and constructing initial cluster by including the seed document and neighbor documents similar to the seed document; (c) making documents, which have document commonality to a current cluster higher than a threshold, belong temporarily to the current cluster; wherein said making comprising: constructing a current cluster common co-occurrence matrix for the current cluster

and a current cluster document frequency matrix of the current cluster based on occurrence frequencies of terms or term pairs appearing in the documents of the current cluster; obtaining a distinctiveness value of each term and each term pair for the current cluster by comparing the input document frequency matrix with the current cluster document frequency matrix; obtaining weights of each term and each term pair from the distinctiveness values; obtaining a document commonality to the current cluster for each document in a input document set based on a product sum between every component of the co-occurrence matrix of the input document and the corresponding component of the current cluster common co-occurrence matrix while applying the weights to said components; and making the documents having document commonality to the current cluster higher than the threshold belong temporarily to the current cluster; (d) repeating step (c) until number of documents temporarily belonging to the current cluster does not increase;", in combination with the other claimed limitations.

With respect to the independent claim 7, the prior art of record, single or in combination, does not teach or fairly suggest the step of: "(a) using a computer to obtain a co-occurrence matrix for the input document, obtaining a co-occurrence matrix S^r for a input document D_r based on occurrence frequencies of terms or term pairs appearing in the set of input documents; wherein in step (a), each mn component S^r_{mn} of the co-occurrence matrix S^r of the document D_r is determined in accordance with: $S^r_{mn} = \sum_{y=1}^{Y_r} d_{ym} d_{yn}$ where: m and n denote m^{th} and n^{th} terms, respectively, among M terms appearing in the set of input documents, D_r is r^{th} document in a document set D consisting of R documents; Y_r is number of document segments

in the document D_r, wherein said d_{rym} and d_{ryn} denote existence or absence of the m^{th} and n^{th} terms, respectively, in y^{th} document segment of the document D_r, and S^r_{mn} represents number of document segments in which the m^{th} term occurs and S^r_{mn} represents co-occurrence counts of document segments in which the m^{th} and n^{th} terms co-occur; (b) selecting a seed document from a set of remaining documents that are not included in any cluster existing, and constructing a current cluster of an initial state based on the seed document, wherein said selecting and said constructing comprise: constructing a remaining document common co-occurrence matrix T^A for a set of the remaining documents based on co-occurrence matrices of all documents in the set of remaining documents; obtaining a document commonality of each remaining document to the set of the remaining documents based on the co-occurrence matrix S^r of each remaining document and the remaining document common co-occurrence matrix T^A; extracting the document having a highest document commonality to the set of the remaining documents; and constructing a initial cluster by including the seed document and neighbor documents similar to the seed document;", in combination with the other claimed limitations.

Any comments considered necessary by applicant must be submitted no later than the payment of the issue fee and, to avoid processing delays, should preferably accompany the issue fee. Such submissions should be clearly labeled "Comments on Statement of Reasons for Allowance".

Contact Information

7. Any inquiry concerning this communication or earlier communications from the examiner should be directed to MICHAEL PHAM whose telephone number is (571)272-3924. The examiner can normally be reached on 9am-5pm.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, John Cottingham can be reached on 571-272-7079. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/M. P./
Examiner, Art Unit 2167

/Luke S. Wassum/
Primary Examiner
Art Unit 2167